T2SGrid: Temporal-to-Spatial Gridification for Video Temporal Grounding

Supplementary Material

1. Training Detail of T2SGrid

All individual video frames are formatted to 336×336. For the Charades-STA dataset [3], both training and evaluation are conducted at 1 fps with the grid configuration set to `g43_s7`, which includes overlap. To reduce computational cost, ActivityNet-Caption [1] is trained and evaluated at 0.5 fps using the `g43_s12` grid configuration, which contains no overlap.

Training is performed using LoRA fine-tuning with `lora_r` = 64 and `lora_alpha` = 128. The model is trained for 3 epochs on $4\times$ NVIDIA A100 40GB GPUs, taking approximately 20 hours. The learning rate is set to 2×10^{-5} with the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$), and a linear scheduler is applied to decay the learning rate.

2. More Experiments

Zero-shot Performance on QVHighlights. We evaluate zero-shot performance on QVHighlights. As shown in Table 1, integrating T2SGrid consistently enhances vision-language models across different scales. Notably, it significantly boosts mAP and HIT@1 for both Qwen2-VL-7B and Qwen3-VL-8B baselines. The Qwen3-VL-8B + T2SGrid combination achieves peak performance (32.5 mAP, 55.3 HIT@1), demonstrating our method's strong zero-shot effectiveness and generalization.

Table 1. Zero-shot performance on QVhighlights

Method	mAP	HIT@1
TimeSuite	26.5	54.1
Trace	26.8	42.7
Qwen2-VL-7B	21.5	42.2
NumPro (based on Qwen2-VL-7B)	23.6	43.4
Ours (Qwen2-VL-7B + T2SGrid)	**24.1**	**44.1**
Qwen3-VL-8B	31.3	51.4
Ours (Qwen3-VL-8B + T2SGrid)	**32.5**	**55.3**

Hybrid Training. We tested a hybrid training approach using both grid and sequential inputs. Incorporating T2SGrid as an auxiliary signal during conventional fine-tuning yielded superior results, boosting the grid-input mIoU to 54.5 and the sequential-input mIoU to a peak of 56.8.

Table 2. Performance comparison of Hybrid Training.

Method	R1@0.3	R1@0.5	R1@0.7	mIoU
FT	73.1	56.8	32.7	50.4
T2SGrid-FT	76.8	60.6	35.9	53.2
FT + T2SGrid-FT (Hybrid Training)				
seq_input	78.7	65.7	42.7	56.8
grid_input	77.7	61.5	37.7	54.5

3. Ablation Study

3.1. Ablation Study on Grid Structure

To validate the necessity of the M×N grid structure, comparative experiments were conducted on the Charades-STA dataset (Qwen2-VL-7B [13]) with three layouts: the standard 4×3 grid (prior ablation studies confirmed its optimality among grid configurations), 12×1 horizontal linear sequence, and 1×12 vertical linear sequence. As shown in Table 3, the grid structure significantly outperforms linear layouts across all metrics. At stride=7, the grid achieves R1@0.3 (70.1), surpassing the horizontal linear (66.8) by 3.3 and the vertical linear (64.9) by 5.2; its R1@0.5 (46.7) exceeds the horizontal linear (41.9) by 4.8 and the vertical linear (41.5) by 5.2. At stride=12, the grid's R1@0.3, R1@0.5, R1@0.7, and mIoU remain universally superior to linear sequences, and results under stride=7 always exceed those under stride=12 under equivalent conditions. These findings demonstrate that the 2D grid structure effectively preserves temporal continuity through compact spatial proximity, whereas the dispersed frame arrangement in linear layouts weakens modeling capability.

Table 3. Grid Structure Performance Comparison (Charades-STA)

Layout Type	Stride	R1@0.3	R1@0.5	R1@0.7	mIoU
T2SGrid(4×3)	7	**70.1**	**46.7**	**20.1**	**44.3**
Horizontal Linear(12×1)	7	66.8	41.9	18.0	42.0
Vertical Linear(1×12)	7	64.9	41.5	17.2	40.6
T2SGrid(4×3)	12	**64.5**	**42.9**	**18.1**	**41.2**
Horizontal Linear(12×1)	12	64.1	42.2	17.5	41.0
Vertical Linear(1×12)	12	63.8	41.7	17.9	40.6

3.2. Ablation Study on Frame Ordering Schemes

Within the 4×3 grid framework, four ordering schemes were evaluated: row-major (left-to-right row-wise traversal), column-major (top-to-bottom column-wise traversal), horizontal snake (left-to-right within rows, with row-wise

direction alternating), and vertical snake (top-to-bottom within columns, with column-wise direction alternating). As presented in Table 4, row-major ordering consistently delivers optimal performance: at stride=7, its R1@0.3 leads column-major (69.0) by 1.1; at stride=12, row-major similarly outperforms column-major and snake arrangements in R1@0.3, R1@0.5, R1@0.7, and mIoU. Critically, all metrics (R1@0.3/R1@0.5/R1@0.7/mIoU) under stride=7 surpass their stride=12 counterparts, further validating the model adapts best to row-major sequencing.

Table 4. Frame Ordering Performance Comparison (Charades-STA)

Ordering Scheme	Stride	R1@0.3	R1@0.5	R1@0.7	mIoU
Row-major	7	**70.1**	**46.7**	**20.1**	**44.3**
Column-major	7	69.0	45.7	19.0	43.5
Horizontal Snake	7	68.8	45.1	18.9	43.2
Vertical Snake	7	68.7	45.8	19.2	43.5
Row-major	12	**64.5**	**42.9**	**18.1**	**41.2**
Column-major	12	64.2	42.5	17.9	41.0
Horizontal Snake	12	63.7	42.7	16.6	38.9
Vertical Snake	12	63.5	42.8	17.0	39.4

3.3. Ablation on Implicit Temporal Encoding in Grid and Explict visualNum

We further conduct an ablation study to examine the implicit temporal encoding mechanism in T2SGrid. While the main text highlights that our method leverages the inherent implicit encoding provided by the grid structure, we additionally explore augmenting each frame within the grid with a visual index placed at the bottom-right corner to provide explicit temporal cues [16]. As shown in Table 5, adding such visual markers leads to performance degradation: R1@0.3 drops by 7.6 and mIoU decreases by 4.2. The extra visual numbering introduces interference to the visual content, causing the model to over-focus on local numeric attributes while overlooking the spatial evolution patterns across frames, thereby weakening its temporal modeling capability. These findings demonstrate that leveraging the intrinsic spatial topology of the grid representation to encode local temporal dependencies is an effective design choice.

Table 5. Ablation Study on Implicit Time Encoding (Charades-STA)

Configuration	R1@0.3	R1@0.5	R1@0.7	mIoU
Implicit	**70.2**	**46.7**	**20.1**	**44.3**
Implicit + visualNum	62.6	41.4	19.1	40.1

4. More Attention Analysis

Qualitative Analysis of Attention. We conducted further attention analysis in Figure 1. From the perspective of spatial attention, sequential-frame inputs exhibit noisy and diffuse attention patterns. For example, in the first and second cases, the attention spreads across irrelevant regions of the frame. In contrast, the grid-based representation enables the model to focus on the woman beginning to eat a sandwich in the first example and the pill-taking action in the second example. This shows that the grid structure effectively enhances local spatial attention, as frames within each grid can attend to one another through the attention mechanism.From the perspective of temporal attention, the grid representation also guides the model to place attention closer to the ground-truth moments, whereas sequential-frame inputs may sometimes deviate from the correct temporal regions.

Quantitative Analysis of Attention. A key limitation of sequential inputs is that the attention weights tend to be uniformly dispersed, struggling to concentrate on the true temporal dynamics. To quantitatively evaluate this, we propose two metrics: **Temporal Attention Entropy (TAE)** and **Temporal Attention Accuracy (TAA)**.

Given the sequence of frame-level attention scores $A = [a_1, a_2, ..., a_T]$ (typically the average of the last cross-attention layer), we first compute the normalized attention distribution $p_t = a_t / \sum_{i=1}^{T} a_i$. The TAE is then defined as:

$$\text{TAE} = -\sum_{t=1}^{T} p_t \log(p_t) \tag{1}$$

Theoretically, sequential inputs yield a higher entropy due to their dispersed attention. In contrast, our proposed grid input (T2SGrid) yields a lower entropy by keeping the attention highly concentrated.

To measure the alignment with the ground truth, we construct a GT mask sequence $G = [g_1, g_2, ..., g_T]$, where $g_t = 1$ if an action occurs at step t, and $g_t = 0$ otherwise. TAA calculates the proportion of the attention mass that correctly falls inside the GT interval:

$$\text{TAA} = \frac{\sum_{t=1}^{T}(a_t \cdot g_t)}{\sum_{t=1}^{T} a_t} \tag{2}$$

As demonstrated in the table below, our T2SGrid achieves notably lower entropy (TAE ↓) and higher accuracy (TAA ↑) compared to the sequential baseline (Seq.). This confirms that T2SGrid successfully yields an attention mechanism that is both highly concentrated and precisely aligned with the target temporal regions.

Input Format	TAE (↓)	TAA (↑)
Sequential (Seq.)	3.229	0.296
T2SGrid (Ours)	**2.781**	**0.458**

5. Adaptability to Long Videos and Varying Frame Rates

Our T2SGrid framework is designed with inherent flexibility to handle videos of varying durations and frame rates without compromising spatial resolution or temporal precision. In this section, we detail how our method T2SGrid adapt to these variations.

5.1. Scalability to Long Videos

Processing long-form videos remains a major challenge for Vision-LLMs [5, 6, 12, 13]. T2SGrid constructs temporally coherent grids through a configurable sliding-window mechanism, enabling the model to scale more robustly to longer video durations.

To illustrate this, consider a 10-minute (600 s) video. When sampled at 1 fps with a temporal stride of $s = k = 12$, the video is divided into roughly $600/12 = 50$ grids, which are processed sequentially by the Video-LLM. Even under a much lower sampling rate such as 0.1 fps, the same video still forms about $60/12 = 5$ grids, demonstrating that T2SGrid naturally adapts to longer videos by adjusting the number of grids.

This flexibility confirms that our design maintains temporal coverage without increasing computational burden, making it well-suited for long-video understanding [10, 11, 14, 15, 17]. As reported in the main text, T2SGrid delivers notable gains on the long-form benchmark VideoMME [2], achieving an improvement of 0.8 points over baseline.

5.2. Robustness to Varying Frame Rates (FPS)

Our method is also inherently robust to varying frame rates, since it operates directly on frame counts rather than absolute timestamps. In principle, this allows T2SGrid to adapt to any FPS without modifying the model architecture or grid configuration. Whether a 12-frame window corresponds to 0.5 seconds at 24 FPS, 2 seconds at 6 FPS, 12 seconds at 1 FPS, or even 24 seconds at 0.5 FPS, the resulting spatial grid representation (G_i) presented to the Vision-LLM remains structurally identical.

This FPS-agnostic design ensures that T2SGrid maintains consistent spatiotemporal encoding across different video sampling rates, enabling stable performance on both high-FPS short clips and low-FPS long videos.

6. More Qualitative Results

6.1. Additional Visualization Results on Charades-STA dataset

Figure 2 visually compares our method with prior state-of-the-art approaches (including TRACE [4] and NumPro [16]) for video temporal grounding (VTG) [1, 3, 7–9] tasks using the Qwen2-VL-7B model on the Charades dataset. Experimental results demonstrate superior performance in action-oriented query scenarios, where videos typically concentrate around 30 seconds. Our method achieves precise temporal localization for both simple actions (e.g., "Person eat the food.") and complex multi-condition queries (e.g., "Person eats sandwich that is seating on side of sink."), with predicted intervals closely aligning with ground truth annotations despite significant deviations in comparative methods. This visualization robustly validates our approach's exceptional precision for action-related queries.

6.2. Additional Visualization Results on ActivityNet dataset

Further in Figure 3, extended visual comparisons on the ActivityNet dataset validate performance differences between our method and competing approaches under Qwen2-VL-7B. This dataset exhibits significant temporal diversity, with specific cases in Figure 3 demonstrating video durations ranging from 30 seconds to 178 seconds, while its queries primarily involve complex event descriptions rather than simple actions. It accurately localizes prolonged events like "A man rides a horse holding a pole and joins other people that play polo.", whereas other methods exhibit temporal misalignments or fragmented coverage (e.g., capturing only initial phases) in longer videos, failing to encompass entire event progressions. These visual proofs substantiate our method's stable temporal localization capabilities for complex events in duration-varying videos.

7. Limitation

While our T2SGrid has proven effective across multiple models and datasets, significantly surpassing previous state-of-the-art models, there are still some limitations:

- Due to computational constraints, all individual video frames are resized to 336×336, and some videos are evaluated at a reduced sampling rate of 0.5 fps. This low frame-rate setting limits the amount of temporal information available to the model, which may lead to a degradation in temporal modeling performance, especially for tasks requiring fine-grained motion understanding.
- Our approach is built on Vision-LLMs with ViT backbones that support native-resolution inputs. For Vision-LLMs that do not yet offer native-resolution processing, directly applying our method may be less suitable, as it relies on assembling raw video frames without losing spatial information. Using models that downscale or do not preserve native resolution could degrade spatial fidelity and undermine this design principle.

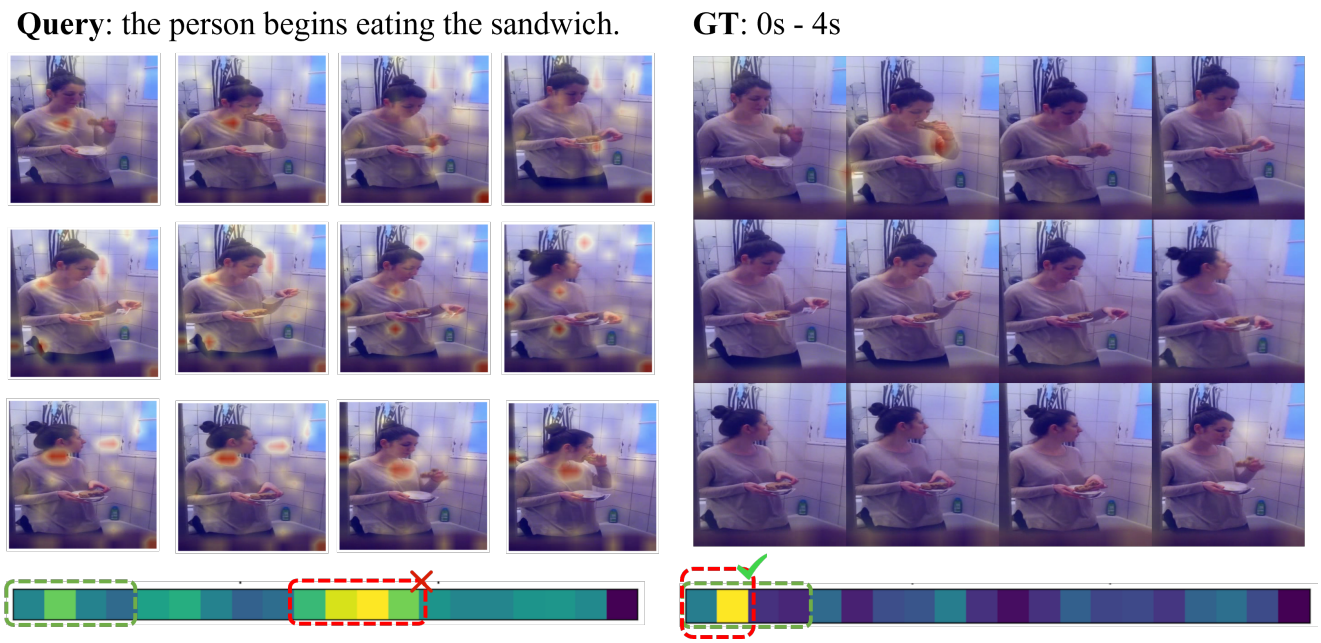

Figure 1. Additional analysis of temporal and spatial attention. Red boxes indicate the model's predictions, while green boxes denote the ground-truth annotations. The left side shows the sequential-frame input, and the right side shows the grid-based input.

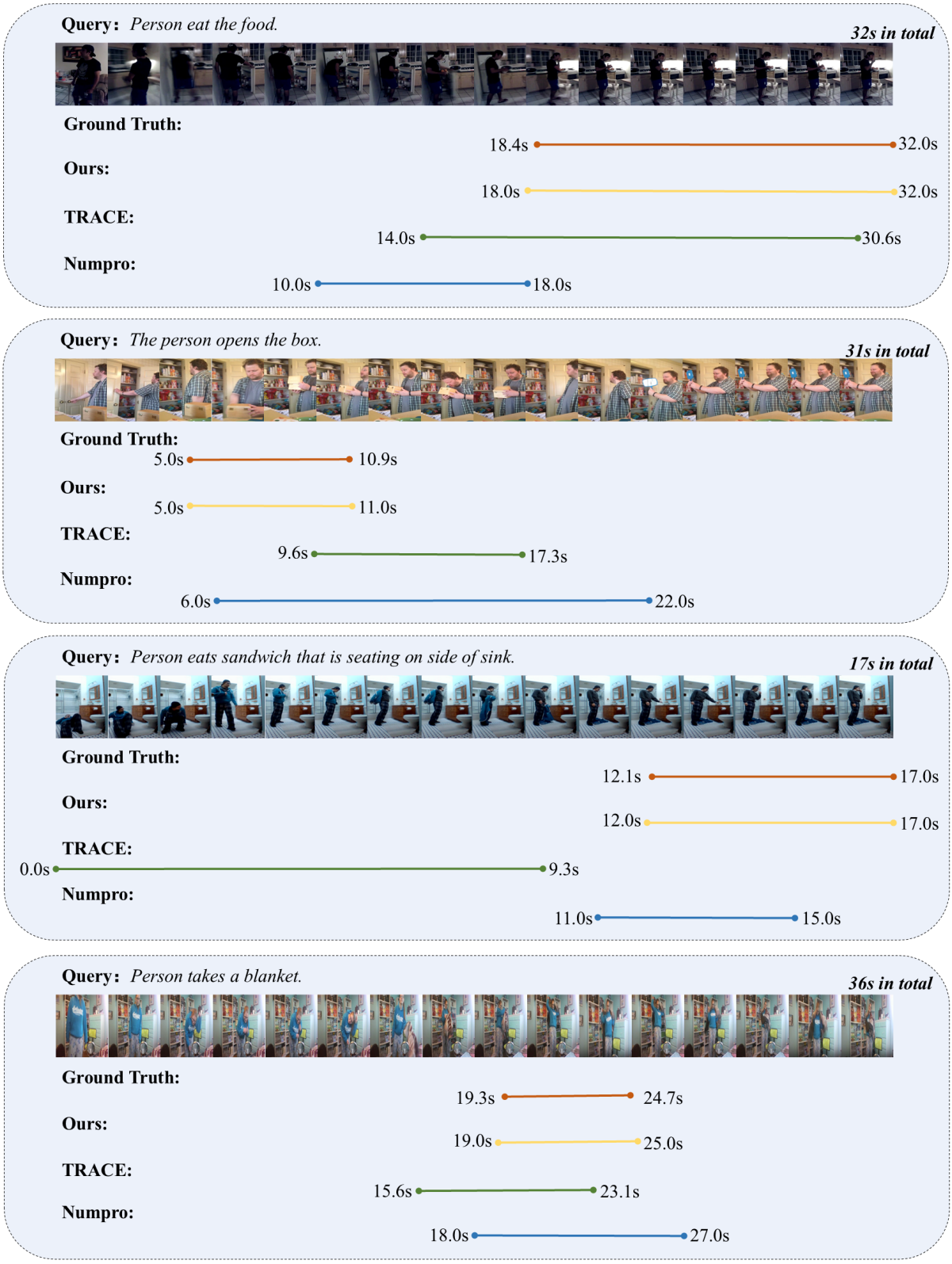

Figure 2. Additional visualization cases of Video Temporal Grounding task on Charades-STA dataset.

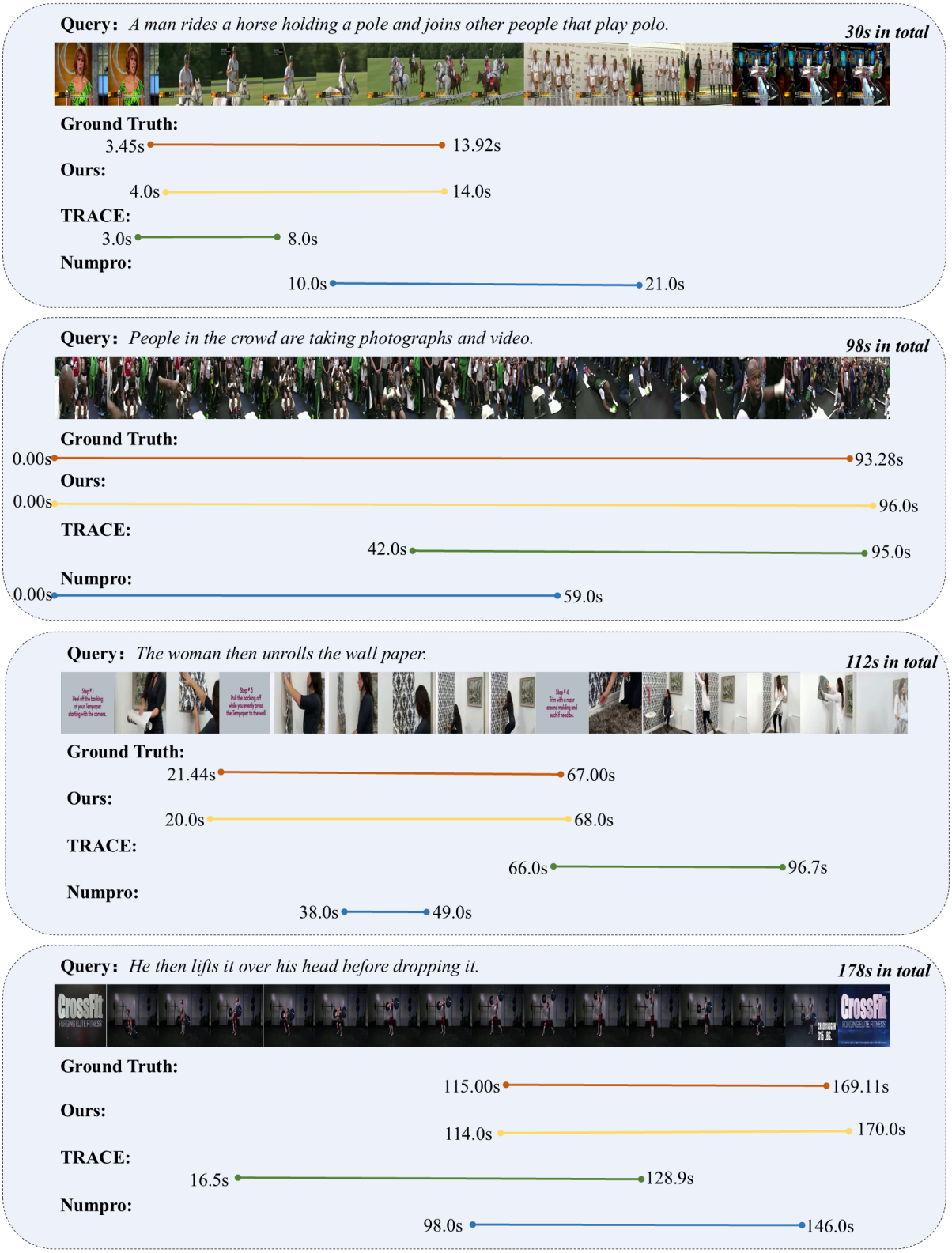

Figure 3. Additional visualization cases of Video Temporal Grounding task on ActivityNet dataset.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1, 3

[2] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 3

[3] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 1, 3

[4] Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024. 3

[5] Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3302–3310, 2025. 3

[6] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

[7] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*, 2024. 3

[8] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*, 2024.

[9] Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. Chatvtg: Video temporal grounding via chat with video dialogue large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1847–1856, 2024. 3

[10] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 3

[11] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[12] Qwen Team. Qwen3-vl: Sharper vision, deeper thought, broader action. *Qwen Blog. Accessed*, pages 10–04, 2025. 3

[13] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 3

[14] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025. 3

[15] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024. 3

[16] Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13754–13765, 2025. 2, 3

[17] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. Timesuite: Improving mllms for long video understanding via grounded tuning, 2025. 3